Designing Your Research Project to Improve Data Quality

Rick White

Department of Statistics, UBC

Graduate Pathways to Success

Faculty of Graduate Studies

October 4, 2011

Discussion Point

What does the word
"Statistics"
mean to you?

The discipline of Statistics

"Statistics is the study of the collection, organization, and interpretation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments." (Wikipedia)

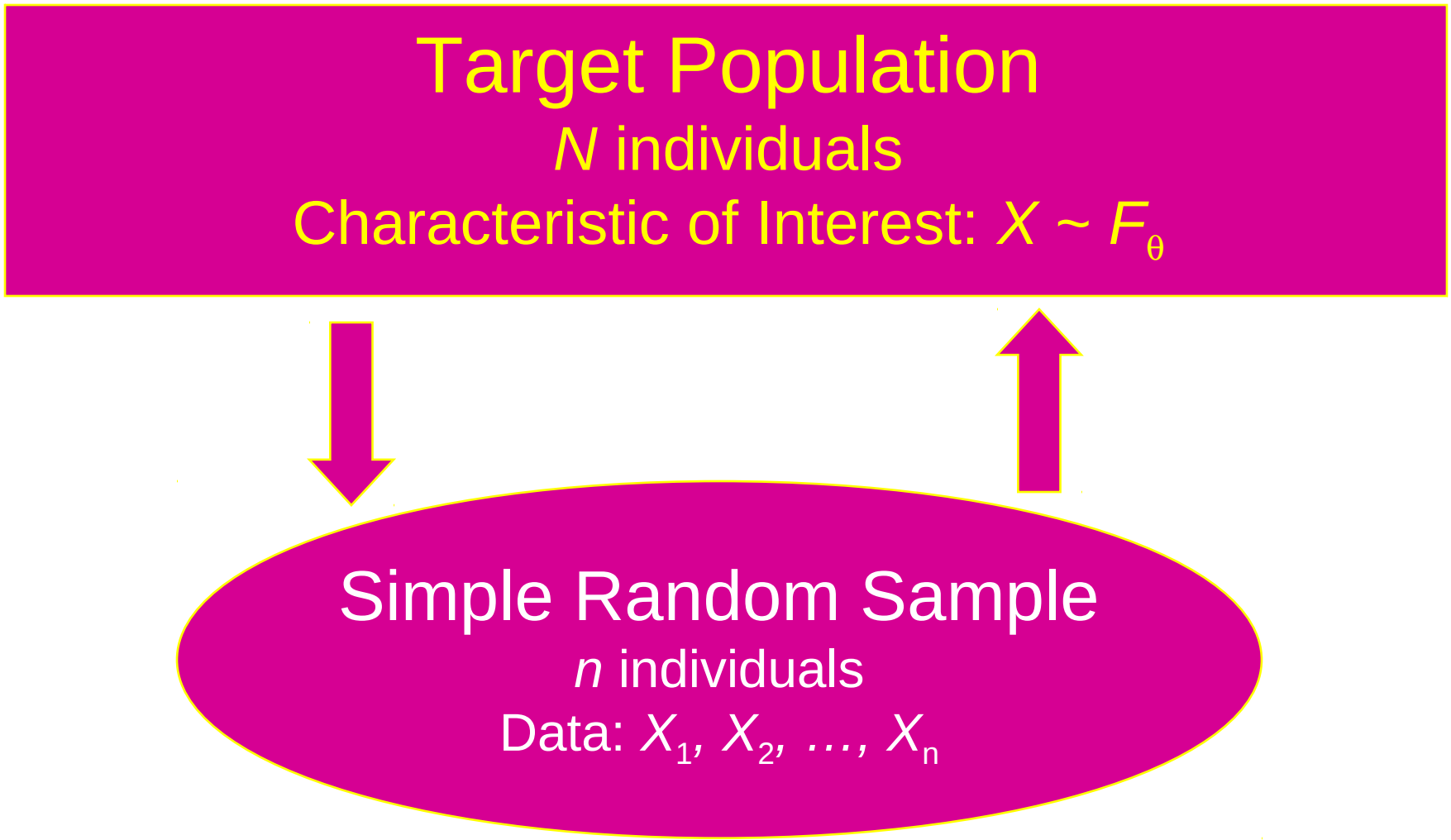
Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers

Types of Research Studies

- Sample Surveys
  - Opinion Polls
  - Market research
- Observational Studies
  - Smoking and cancer studies
- Controlled Experiments
  - Clinical trials
  - Quality control methods

Sample Surveys: Basic Concepts



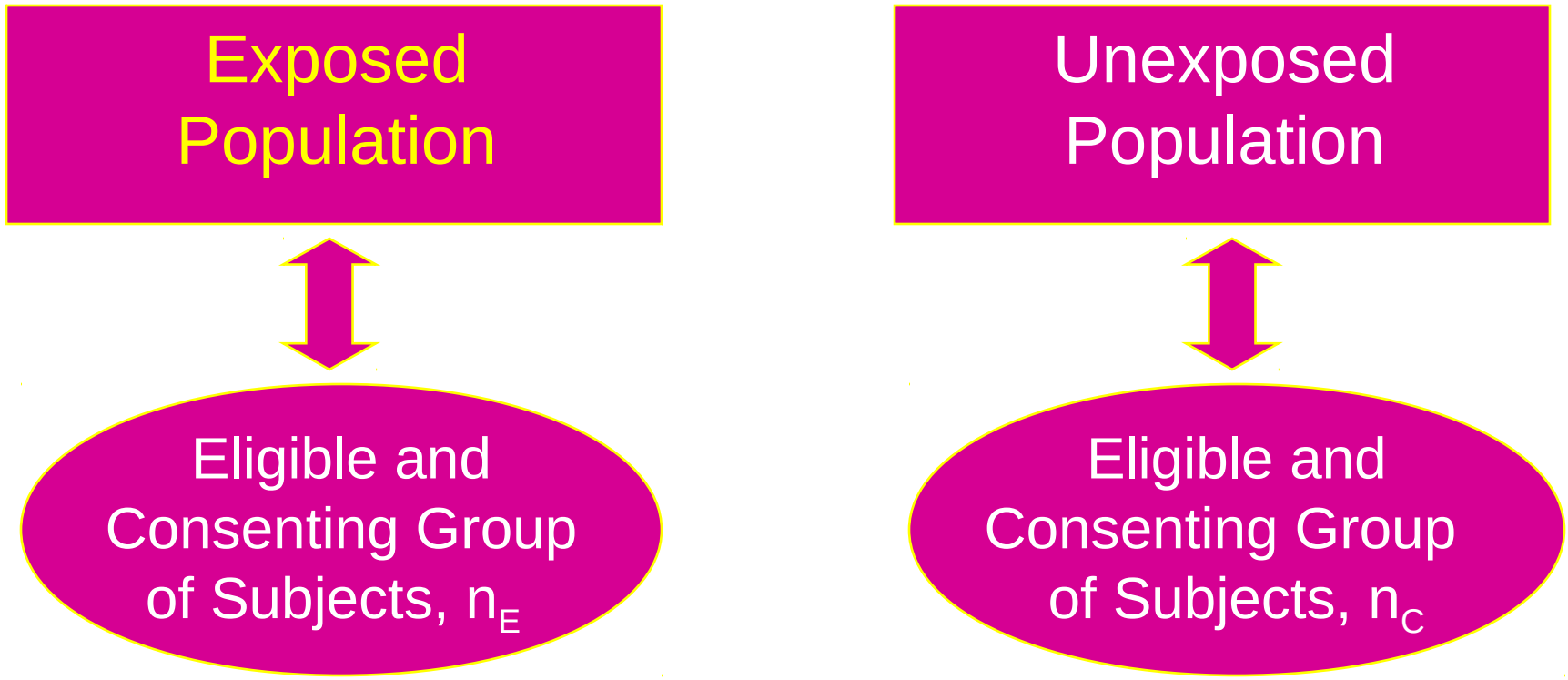
Sample Surveys: Some Issues

- Establishing the sampling frame.
- Over-coverage or Under-coverage.

  Is the entire population in the sampling frame?

  Are elements not in the population in the frame?
- Selection bias.
  - Is the sample *representative* of the target population?
- Non-response bias.
  - Do non-responders *differ* from responders?
- Wording and order of the questions

A Two-Group Observational Study



Observational Studies: Some Issues

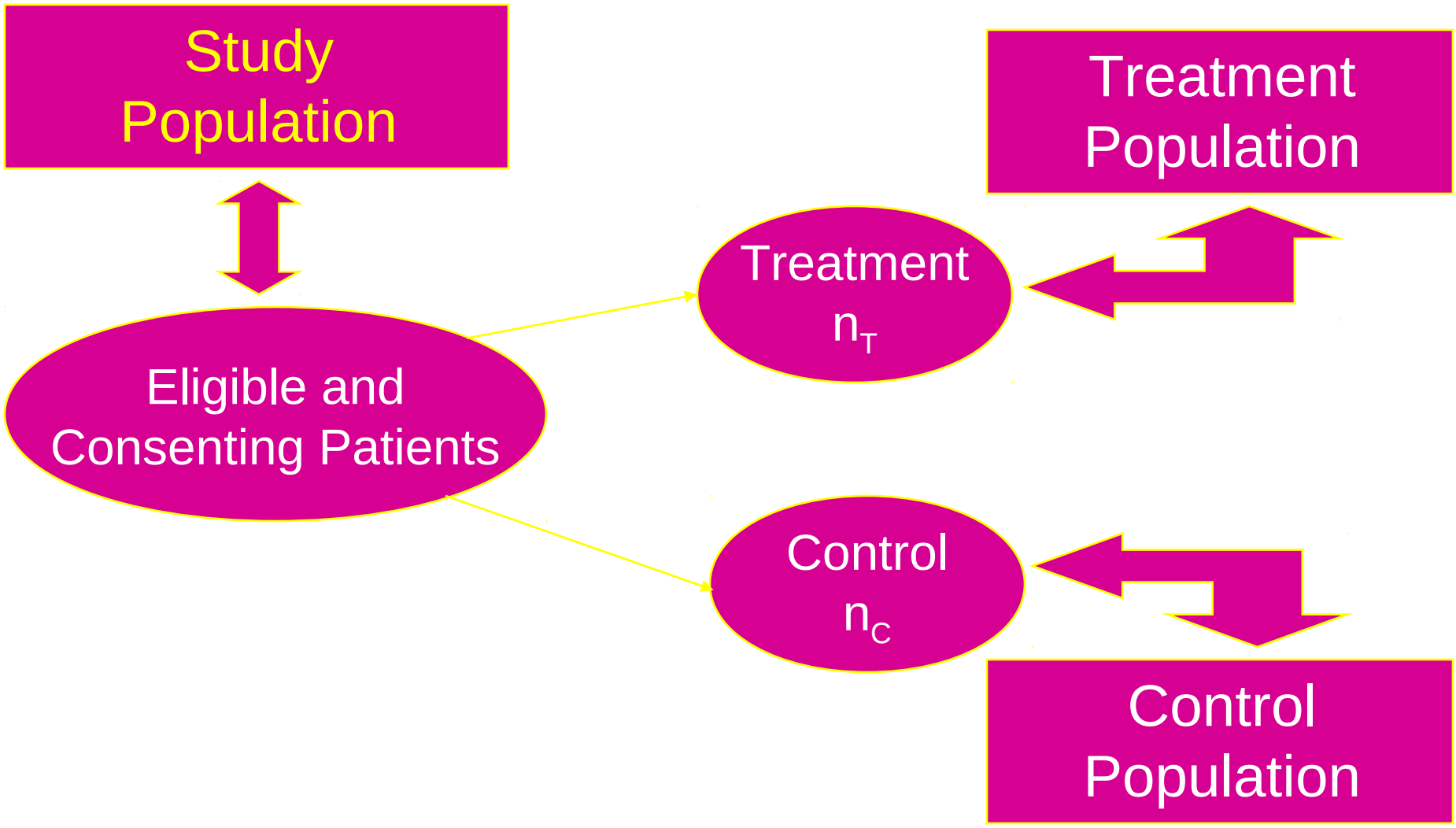
- Describing differences is straightforward
  - Examination of associations with exposure
- Confounding is a major issue
  - Differences could be due to a factor other than exposure that differs across the groups
  - Analyses must adjust for such confounders
- Analyses can suggest, but not establish causality
  - Bradford Hill's criteria for causality
- If based on samples, are samples representative?

Confounding Example

- Civil Rights Act of 1964.

	Democrat	Republican
Northern State	**94% (145/154)**	85% (138/162)
Southern State	**7% (7/94)**	0% (0/10)
Combined	61% (152/248)	**82% (138/172)**

A Two-Armed Clinical Trial



Clinical Trials: Some Issues

- Ethics of experimentation.
  - Cannot force someone to smoke
- Choice of control: nothing, placebo or active.
  - The placebo effect
  - ethically must treat if one already exists
- Blinding of subjects and evaluators.
  - subconscious effects
- Target population versus study population.

Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers

The Population

- This is who or what you are going to study.
- Observational studies may compare more than 1 population.
- Controlled experiments usually apply treatments to the population.
- Populations are hard to deal with in surveys without a proper sampling frame.

Defining the Question

- Make sure questions are clear and focused.
- All questions should be based on the same populations or perhaps a subset
- Each question should define a single hypothesis to test or quantity to measure.
- “Does Betaseron decrease the relapse rate in relapsing-remitting MS patients?”

The Statistical Hypothesis

- Null Hypothesis: the default condition
  - Absence of evidence is not evidence of absence
- Alternative Hypothesis: what we want to show
- α = significance level. It is the chance of a false positive (type 1 error).
  - Does not depend on the sample size.
- β = power. It is the chance of a true positive.
  - It is a function of the sample size
- Select sample size to give desired power.

Hypothesis test

- Define null and alternative hypothesis
- Determine threshold needed to reject null
- Collect data needed to test hypothesis
- Analyze data to determine result
- Discussion point: How is a courtroom trial like a hypothesis test?
    - What part of the trial plays the role of the hypothesis, data collection, the data itself, the data analysis and the result?

Courtroom Trial Analogy

Defendant is innocent until proven guilty

	Null Hypothesis is true Did not commit Crime	Alt. Hypothesis is true Committed Crime
Do not reject Null Hypothesis Not Guilty	Right Decision	Wrong Decision (Type 2 error)
Reject Null Hypothesis Guilty	Wrong Decision (Type 1 error)	Right Decision

Multiple Questions

- The more questions the greater the overall type 1 error rate (familywise error rate).
- Worse case scenario: Independent events
  - P(A & B) = P(A)*P(B)
- 10 hypotheses with $\alpha = 0.05$
  - probability at least 1 false positive = 0.40
- Bonferroni correction $\alpha^* = \alpha/n$
- 10 hypotheses with $\alpha^* = 0.05/10 = 0.005$
  - probability at least 1 false positive = 0.049

Multiple Question: cont

- Hypotheses in same study are typically not independent
- Bonferroni is very conservative and leads to very large sample sizes
- Other less conservative methods are available such as False Discovery Rate
- Recommendations:
  - limit the number of questions
  - select a few questions as primary

Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers

The Sampling Plan

- Very important for surveys.
- Usually based on available population for experiments and observational studies
- Randomization for controlled experiments
- Does the sample represent the population
  - This is who is sampled not how many.
- Does the sample give enough precision?
  - This is how many are sampled.

Random Sampling

- Simple Random Sampling (lottery)
  - Each member of the population has an equal chance of being sampled.
  - Requires a complete sampling frame
- Systematic Random Sampling
  - requires an ordered sampling frame, a random start point and a sampling interval.
- Stratified Sampling
- Cluster Sampling

Non Random Sampling

- Convenience or Accidental sampling
  - sample people who pass by
  - using an online survey tool
  - probably the most common type of sample
- Judgement sampling
  - The researcher chooses who to sample
- Snowball sampling
  - respondents get their friends to respond

Roosevelt vs Landon: 1936

- Election Poll by Literary Digest
  - 10 million sampled from their readers, telephone users, and automobile owners
  - n = 2.4 million responded (huge sample size)
  - result: 57% for Landon, 36% for Roosevelt
- Actual Election Result
  - 36.5% for Landon, 60.8% for Roosevelt
- Discussion point: Why do you think Literary Digest was so inaccurate?

Gallup Polls

- George Gallup correctly predicted Roosevelt vs. Landon with n = 5000
- Gallup also sampled 50000 and correctly predicted Literary Digest result to within 1%.
- Gallup Polls are used in more than 140 countries today
- Gallup Polls are not always correct.
    - Dewey vs Truman 1948
    - Ford vs Carter 1976

Sample Properties

- Random Samples
  - representative of the population
  - have a measurable sampling error
  - non-response can introduce bias
  - incomplete frame can introduce coverage bias
- Non random samples
  - Relationship to target population is immeasurable
  - bias in sample is unknown
  - results cannot be applied to the population

Randomization

- Usually done by random number generators and permutation algorithms on computers
- Should equally assign a unit to any treatment or control group based on relative sample size.
- A form of simple random sampling without replacement applied to the population sample.
- Eliminates bias from the study. Bias is defined in terms of expectation not result.

Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers

The Data

- Does it allow you to answer your questions?
- Data quality issues
  - Was it collected in a reliable and accurate fashion?
  - Was it recorded in a reliable and accurate fashion?
  - Is it complete or are several pieces of information missing?
- The data should be locked before analysis.
- Discussion Point: Why should the data be finalized before analysis begins?

Types of variables

- Numeric variables
  - time to event, number of successes
  - can be continuous or discrete
- Categorical variables
  - ordinal (small, medium, large)
  - nominal (red, green, blue)
- Response variables (dependent variables)
- Explanatory variables (independent variables)

Experimental Unit

- The level of randomization or sampling.
- Sampling unit does not always equal the experimental unit
- Independent observations.
- Pseudoreplication are observations within an experimental unit: repeated measures
  - results in over exaggeration of statistical significance

Types of Analyses

- Univariate analysis:
  - only 1 response variable in the model
  - may have many explanatory variables in the model
- Multivariate analysis
  - more than 1 response variables in the same model
- Some people think multivariate models are 1 response with many explanatory variables
- Make sure you are talking about the same thing.

Common Analyses

- Continuous response
  - t-test, paired t-test, rank tests
  - regression, ANOVA, linear models
- Count or Binary response:
  - Poisson regression (over-dispersion)
  - contingency tables, logistic regression
- Categorical responses are more complicated
- Pseudoreplication complicates the analysis
  - Mixed effects modeling

Sample Size Considerations

- Most statistical analysis depend on a reasonable sample size to be valid
- As sample size increases
  - statistical power and precision increases
  - statistical accuracy does not change
  - distribution of statistics become more predictable
- Demonstration http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers

Planning your Study

- Define the questions of interest.
- Determine the appropriate populations that will allow the questions to be answered.
- Create a plan to sample the populations. Randomization may be required.
- Determine what information is needed from the sample to answer the questions.
- Create an appropriate analysis plan.

Statisticians can help

- Focus and clarify the objectives
- Design an appropriate sampling plan
- Provide a randomization scheme
- Design an appropriate analysis plan

➔ Talk to a statistician before you collect data !!!

Statistical Resources

- Statistical Consulting and Research Laboratory (SCARL)

  www.stat.ubc.ca/SCARL

- Stat551 practicum course
- Short Term Consulting Service
- SCARL staff hourly Service

Outline

- Main Types of Studies
- Defining your Questions
- Sampling and Randomization
- The Data
- Putting it together
- Questions and Answers